

An Improved Composite Hypothesis Test for Markov Models with Applications in Network Anomaly Detection*

Jing Zhang[†] and Ioannis Ch. Paschalidis[‡]

Abstract—Recent work has proposed the use of a composite hypothesis Hoeffding test for statistical anomaly detection. Setting an appropriate threshold for the test given a desired false alarm probability involves approximating the false alarm probability. To that end, a large deviations asymptotic is typically used which, however, often results in an inaccurate setting of the threshold, especially for relatively small sample sizes. This, in turn, results in an anomaly detection test that does not control well for false alarms. In this paper, we develop a tighter approximation using the Central Limit Theorem (CLT) under Markovian assumptions. We apply our result to a network anomaly detection application and demonstrate its advantages over earlier work.

Index Terms—Hoeffding test, weak convergence, Markov chains, network anomaly detection.

I. INTRODUCTION

During the last decade, the applications of statistical anomaly detection in communication networks using the Large Deviations Theory (LDT) [1] have been extensively explored; see, e.g., [2], [3], [4], among others. Statistical anomaly detection involves characterizing the “normal behavior” of the system and identifying the time instances corresponding to abnormal system behavior. Assuming that the network traffic is stationary in time, [2] applies two methods to that end. The first method, termed “*model-free*,” models traffic as an independent and identically distributed (i.i.d.) sequence. The second method, termed “*model-based*,” models traffic as a finite-state Markov chain. These are then extended in [4] to the case where the network traffic is time-varying. Essentially, each of these methods is designed to tackle a certain Universal Hypothesis Testing Problem (UHTP).

A UHTP aims to test the hypothesis that a given sequence of observations is drawn from a known *Probability Law (PL)* (i.e., *probability distribution*) defined on a finite alphabet [5]. It is well known that the test proposed by Hoeffding [6] is optimal in an error exponent sense [1], [2], [4]. When implementing Hoeffding’s test, one must set a threshold η , which can be estimated by using Sanov’s theorem [1], a result within the scope of LDT. Note that such an estimate (let us denote it by η^{sv}) is valid only in the asymptotic sense.

In practice, however, only a finite number of observations are available, and it can be shown by experiments (e.g., using the software package TAHTIID [7]) that η^{sv} is typically not accurate enough.

To improve the accuracy of the estimation for η , [5] (or, see [8]) borrows a method typically used by statisticians; that is, deriving results based on Weak Convergence (WC) of the test statistic to approximate the error probabilities of Hoeffding’s test. Under the i.i.d. assumption, by using the technique combining WC results together with Taylor’s series expansion, [5] (or, see [8], [9]) gives an alternative approximation for η (let us denote it by η^{wc}). Using [7], one can verify that, in the finite sample-size setting, η^{wc} is typically much more accurate than η^{sv} .

It is worth pointing out that some researchers have also tried to obtain a more accurate approximation for η by refining Sanov’s theorem [10]. However, as noted in [8], such refinements of large deviation results are typically hard to calculate.

In this work we establish η^{wc} under a Markovian assumption, thus, extending the work of [5] which considered the i.i.d. case. We apply the proposed procedure to network anomaly detection by embedding it into the software package SADIT [11].

The rest of the paper is organized as follows. We first formulate the problem in Sec. II and derive the theoretical results in Sec. III. We then present the experimental results in Sec. IV and finally provide concluding remarks in Sec. V.

Notation: By convention, all vectors are column vectors. To save space, we write $\mathbf{x} = (x_1, \dots, x_n)$ to denote the column vector \mathbf{x} . We use prime to denote the transpose of a matrix or vector.

II. PROBLEM FORMULATION

Let $\Xi = \{\xi_i; i = 1, \dots, N\}$ be a finite alphabet. The observations $\mathbf{Y} = \{Y_l; l = 0, 1, 2, \dots\}$ form a stochastic process taking values from Ξ .

In the *model-based* UHTP [1], [2], [4], under the null hypothesis \mathcal{H} , the observations $\mathbf{Y} = \{Y_l; l = 0, 1, 2, \dots\}$ are drawn according to a Markov chain with state set Ξ . Let its transition matrix be $\mathbf{Q} = (q_{ij})_{i,j=1}^N$.

Let $\mathbf{1}_{\{\cdot\}}$ denote the indicator function. Define the *empirical PL* by

$$\Gamma_n(\theta_{ij}) = \frac{1}{n} \sum_{l=1}^n \mathbf{1}_{\{Z_l = \theta_{ij}\}}, \quad (1)$$

where $Z_l = (Y_{l-1}, Y_l)$, $l = 1, \dots, n$, $\theta_{ij} = (\xi_i, \xi_j) \in \Xi \times \Xi$, $i, j = 1, \dots, N$, and denote the new alphabet $\Theta =$

* Research partially supported by the NSF under grants CNS-1239021 and IIS-1237022, by the ARO under grants W911NF-11-1-0227 and W911NF-12-1-0390, by the ONR under grant N00014-10-1-0952, and by the Cyprus Research Promotion Foundation under Grant New Infrastructure Project/Strategic/0308/26.

[†] Division of Systems Engineering, Boston University, jzh@bu.edu.
[‡] Department of Electrical and Computer Engineering and Division of Systems Engineering, Boston University, 8 St. Mary’s St., Boston, MA 02215, yannisp@bu.edu, <http://sites.bu.edu/paschalidis>.

$\{\theta_{ij}; i, j = 1, \dots, N\} = \{\tilde{\theta}_k; k = 1, \dots, N^2\}$. Note that $\Theta = \Xi \times \Xi$ and $\tilde{\theta}_1 = \theta_{11}, \dots, \tilde{\theta}_N = \theta_{1N}, \dots, \tilde{\theta}_{(N-1)N+1} = \theta_{N1}, \dots, \tilde{\theta}_{N^2} = \theta_{NN}$. Let also the set of PLs on Θ be $\mathcal{P}(\Theta)$.

It is seen that the transformed observations $\mathbf{Z} = \{Z_l; l = 1, 2, \dots\}$ form a Markov chain evolving on Θ . Let its transition matrix be $\mathbf{P} = (p_{ij})_{i,j=1}^{N^2}$ with stationary distribution

$$\boldsymbol{\pi} = (\pi_{ij}; i, j = 1, \dots, N) = (\tilde{\pi}_k; k = 1, \dots, N^2),$$

where π_{ij} denotes the probability of seeing θ_{ij} , and $\tilde{\pi}_1 = \pi_{11}, \dots, \tilde{\pi}_N = \pi_{1N}, \dots, \tilde{\pi}_{(N-1)N+1} = \pi_{N1}, \dots, \tilde{\pi}_{N^2} = \pi_{NN}$. Let $p(\theta_{ij}|\theta_{kl})$ denote the transition probability from θ_{kl} to θ_{ij} . Then we have

$$p(\theta_{ij}|\theta_{kl}) = \mathbf{1}_{\{i=l\}} q_{ij}, \quad k, l, i, j = 1, \dots, N,$$

which enables us to obtain \mathbf{P} directly from \mathbf{Q} .

Write $\boldsymbol{\pi} = (\pi_{11}, \dots, \pi_{1N}, \dots, \pi_{N1}, \dots, \pi_{NN})$, and $\boldsymbol{\Gamma}_n = (\Gamma_n(\theta_{11}), \dots, \Gamma_n(\theta_{1N}), \dots, \Gamma_n(\theta_{N1}), \dots, \Gamma_n(\theta_{NN}))$.

Define the *relative entropy* (i.e., *divergence*)

$$D(\boldsymbol{\Gamma}_n \| \boldsymbol{\pi}) = \sum_{i=1}^N \sum_{j=1}^N \Gamma_n(\theta_{ij}) \log \frac{\Gamma_n(\theta_{ij}) / \sum_{j=1}^N \Gamma_n(\theta_{ij})}{\pi_{ij} / \sum_{j=1}^N \pi_{ij}}, \quad (2)$$

and the *empirical measure*

$$\mathbf{U}_n = \sqrt{n}(\boldsymbol{\Gamma}_n - \boldsymbol{\pi}). \quad (3)$$

For any positive integer n , let \mathcal{H}_n be the output of the test that decides to accept or to reject the null hypothesis \mathcal{H} based on the first n observations in the sequence \mathbf{Z} . Under the Markovian assumption, Hoeffding's test [1] is given by

$$\mathcal{H}_n = \mathcal{H} \Leftrightarrow D(\boldsymbol{\Gamma}_n \| \boldsymbol{\pi}) \leq \eta_n, \quad (4)$$

where $D(\boldsymbol{\Gamma}_n \| \boldsymbol{\pi})$, defined in (2), is the *test statistic*, and η_n is a *threshold*. Note that, when applied in anomaly detection, Hoeffding's test will report an anomaly if and only if $\mathcal{H}_n \neq \mathcal{H}$, or, equivalently, $D(\boldsymbol{\Gamma}_n \| \boldsymbol{\pi}) > \eta_n$.

Under hypothesis \mathcal{H} , the *false alarm probability* [5] of the test sequence $\{\mathcal{H}_n\}$ as a function of n is defined by

$$\beta_n = P_{\mathcal{H}}\{D(\boldsymbol{\Gamma}_n \| \boldsymbol{\pi}) > \eta_n\}, \quad (5)$$

where $P_{\mathcal{H}}\{A\}$ denotes the probability of event A under hypothesis \mathcal{H} . Sanov's theorem [1] implies the approximation for η_n

$$\eta_n^{sv} \approx -\frac{1}{n} \log(\beta_n). \quad (6)$$

Given $\beta_n \in (0, 1)$, our central goal in this paper is to improve the accuracy of the approximation for η_n .

III. THEORETICAL RESULTS

We introduce the following condition:

Condition (C). $\mathbf{Z} = \{Z_l; l = 1, 2, \dots\}$ is an aperiodic, irreducible, and positive recurrent Markov chain [12], [13] evolving on Θ with stationary distribution $\boldsymbol{\pi}$, and with the same $\boldsymbol{\pi}$ as its initial distribution.

Remark 1: Since Θ is a finite set, we know that \mathbf{Z} is uniformly ergodic [12], [14]. Assuming the initial distribution

to be $\boldsymbol{\pi}$ is for notational simplicity; our results apply for any feasible initial distribution. Note also that, under (C), $\boldsymbol{\pi}$ must have full support over Θ , meaning each entry in $\boldsymbol{\pi}$ is strictly positive.

Lemma III.1 Assume (C) holds. Then

$$\frac{\pi_{ij}}{\sum_{t=1}^N \pi_{it}} = \frac{\pi_{ij}}{\sum_{t=1}^N \pi_{ti}} = q_{ij}, \quad i, j = 1, \dots, N. \quad (7)$$

Proof: Expanding the first N entries of $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$, we obtain $q_{1i} \sum_{t=1}^N \pi_{t1} = \pi_{1i}$, $i = 1, \dots, N$. Summing up both sides of these equations, it follows

$$\left(\sum_{i=1}^N q_{1i}\right) \cdot \sum_{t=1}^N \pi_{t1} = \sum_{t=1}^N \pi_{t1}. \quad (8)$$

Noticing $\sum_{i=1}^N q_{1i} = 1$, we see that (8) implies $\sum_{t=1}^N \pi_{t1} = \sum_{t=1}^N \pi_{1t}$, which, together with $q_{11} \sum_{t=1}^N \pi_{t1} = \pi_{11}$, yields

$$\frac{\pi_{11}}{\sum_{t=1}^N \pi_{1t}} = \frac{\pi_{11}}{\sum_{t=1}^N \pi_{t1}} = q_{11}.$$

Similarly, we can show (7) holds for all the other (i, j) 's. ■

Remark 2: By Remark 1 and Lemma III.1 we see that, under Condition (C), all entries in \mathbf{Q} are strictly positive, indicating that any two states of the original chain \mathbf{Y} are connected. This is a strict condition; yet, in engineering practice, if some π_{ij} in (7) are equal to zero, we can always replace them with a small positive value, and then normalize the modified vector $\boldsymbol{\pi}$, thus ensuring that Condition (C) can be satisfied. See Sec. IV-A for an illustration of this trick.

A. Weak convergence of \mathbf{U}_n

Let us first establish CLT results for one-dimensional empirical measures

$$U_n^{(k)} = \sqrt{n}(\Gamma_n(\tilde{\theta}_k) - \tilde{\pi}_k), \quad k = 1, \dots, N^2. \quad (9)$$

Fix $k \in \{1, \dots, N^2\}$. Define

$$f_k(Z) = \mathbf{1}_{\{Z=\tilde{\theta}_k\}} = \begin{cases} 1, & \text{if } Z = \tilde{\theta}_k, \\ 0, & \text{if } Z \in \Theta \setminus \{\tilde{\theta}_k\}. \end{cases} \quad (10)$$

Lemma III.2 Assume (C) holds. Then a CLT holds for $U_n^{(k)}$; that is, $U_n^{(k)} \xrightarrow[n \rightarrow \infty]{d.} \mathcal{N}(0, \sigma_k^2)$ with $\sigma_k^2 = \text{Cov}(f_k(Z_1), f_k(Z_1)) + 2 \sum_{m=1}^{\infty} \text{Cov}(f_k(Z_1), f_k(Z_{1+m})) < \infty$.

Proof: This can be established by applying [12, Corollary 1]. Noting $f_k(\cdot)$ is bounded and the chain \mathbf{Z} is uniformly ergodic, we see that all the conditions needed by [12, Corollary 1] are satisfied. ■

Now we can directly extend Lemma III.2 to the multidimensional case (for an informal argument, see, e.g., [15]; cf. [16, Sec. III.6] for the definition of a Gaussian random variable with a general normal distribution). Due to its role in the following applications, we state the result as a theorem.

Theorem III.3 Assume (C) holds. Then a multidimensional CLT holds for \mathbf{U}_n ; that is,

$$\mathbf{U}_n \xrightarrow[n \rightarrow \infty]{d.} \mathcal{N}(0, \boldsymbol{\Lambda}) \quad (11)$$

with $\mathbf{\Lambda}$ being an $N^2 \times N^2$ covariance matrix given by

$$\mathbf{\Lambda} = \mathbf{\Lambda}_0 + \sum_{m=1}^{\infty} \mathbf{\Lambda}_m, \quad (12)$$

where $\mathbf{\Lambda}_0$ and $\mathbf{\Lambda}_m$ are specified, respectively, by $\mathbf{\Lambda}_0 = [\text{Cov}(f_i(Z_1), f_j(Z_1))]_{i,j=1}^{N^2}$ and $\mathbf{\Lambda}_m = [\text{Cov}(f_i(Z_1), f_j(Z_{1+m})) + \text{Cov}(f_j(Z_1), f_i(Z_{1+m}))]_{i,j=1}^{N^2}$, $m = 1, 2, \dots$

The rest of this subsection is devoted to the computation of $\mathbf{\Lambda}$; we will express $\mathbf{\Lambda} = [\Lambda^{(i,j)}]_{i,j=1}^{N^2}$ in terms of the quantities determining the probabilistic structure of the chain \mathbf{Z} .

In particular, we have

$$\Lambda_0^{(i,j)} = \text{Cov}(\mathbf{1}_{\{Z_1=\tilde{\theta}_i\}}, \mathbf{1}_{\{Z_1=\tilde{\theta}_j\}}), \quad i, j = 1, \dots, N^2, \quad (13)$$

and

$$\begin{aligned} \Lambda_m^{(i,j)} &= \text{Cov}(\mathbf{1}_{\{Z_1=\tilde{\theta}_i\}}, \mathbf{1}_{\{Z_{1+m}=\tilde{\theta}_j\}}) \\ &\quad + \text{Cov}(\mathbf{1}_{\{Z_1=\tilde{\theta}_j\}}, \mathbf{1}_{\{Z_{1+m}=\tilde{\theta}_i\}}), \end{aligned} \quad (14)$$

$m = 1, 2, \dots; i, j = 1, \dots, N^2$.

We first determine (13) as follows. By direct calculation, we obtain $\Lambda_0^{(1,1)} = \tilde{\pi}_1(1 - \tilde{\pi}_1)$, $\Lambda_0^{(1,2)} = -\tilde{\pi}_1\tilde{\pi}_2$, and so on. In general, we have $\Lambda_0^{(i,j)} = \tilde{\pi}_i(\mathbf{I}_{ij} - \tilde{\pi}_j)$, $i, j = 1, \dots, N^2$, where \mathbf{I}_{ij} denotes the (i, j) -th entry of the identity matrix.

Now we compute $\Lambda_2^{(i,j)}$ by (14). Omitting the details, again by direct calculation, we arrive at $\text{Cov}(\mathbf{1}_{\{Z_1=\tilde{\theta}_i\}}, \mathbf{1}_{\{Z_3=\tilde{\theta}_j\}}) = \tilde{\pi}_i(\mathbf{P}_{ij}^2 - \tilde{\pi}_j)$, where \mathbf{P}_{ij}^2 is the (i, j) -th entry of the matrix \mathbf{P}^2 (the square of the transition matrix \mathbf{P}). Interchanging the indexes i, j , we obtain $\text{Cov}(\mathbf{1}_{\{Z_1=\tilde{\theta}_j\}}, \mathbf{1}_{\{Z_3=\tilde{\theta}_i\}}) = \tilde{\pi}_j(\mathbf{P}_{ji}^2 - \tilde{\pi}_i)$. Thus, we have $\Lambda_2^{(i,j)} = \tilde{\pi}_i(\mathbf{P}_{ij}^2 - \tilde{\pi}_j) + \tilde{\pi}_j(\mathbf{P}_{ji}^2 - \tilde{\pi}_i)$.

Computing in a similar way, in general, we derive

$$\Lambda_m^{(i,j)} = \tilde{\pi}_i(\mathbf{P}_{ij}^m - \tilde{\pi}_j) + \tilde{\pi}_j(\mathbf{P}_{ji}^m - \tilde{\pi}_i), \quad m = 1, 2, \dots,$$

where \mathbf{P}_{ij}^m is the (i, j) -th entry of the matrix \mathbf{P}^m (the m -th power of the transition matrix \mathbf{P}).

Thus, for $i, j = 1, \dots, N^2$ we finally attain

$$\Lambda^{(i,j)} = \tilde{\pi}_i(\mathbf{I}_{ij} - \tilde{\pi}_j) + \sum_{m=1}^{\infty} [\tilde{\pi}_i(\mathbf{P}_{ij}^m - \tilde{\pi}_j) + \tilde{\pi}_j(\mathbf{P}_{ji}^m - \tilde{\pi}_i)]. \quad (15)$$

B. Taylor's series expansion of $D(\mathbf{\Gamma}_n \| \boldsymbol{\pi})$

Inspired by [5] (or, see [8]) wherein the i.i.d. assumption is required, we use a Taylor's series expansion to approximate the relative entropy $D(\mathbf{\Gamma}_n \| \boldsymbol{\pi})$.

To this end, for $\mathbf{v} \in \mathcal{P}(\Theta)$ let us consider

$$h(\mathbf{v}) = D(\mathbf{v} \| \boldsymbol{\pi}) = \sum_{i=1}^N \sum_{j=1}^N v_{ij} \log \frac{\frac{v_{ij}}{\sum_{t=1}^N v_{it}}}{\frac{\pi_{ij}}{\sum_{t=1}^N \pi_{it}}}. \quad (16)$$

By direct calculation, we derive

$$\begin{aligned} \frac{\partial h(\mathbf{v})}{\partial v_{ij}} &= \log v_{ij} - \log \left(\sum_{t=1}^N v_{it} \right) - \log \pi_{ij} \\ &\quad + \log \left(\sum_{t=1}^N \pi_{it} \right), \quad i, j = 1, \dots, N, \end{aligned} \quad (17)$$

which leads to

$$\nabla h(\boldsymbol{\pi}) = 0. \quad (18)$$

Further, from (17), we compute the Hessian $\nabla^2 h(\mathbf{v})$ by considering three cases:

(i) If $k \neq i$, then $\frac{\partial^2 h(\mathbf{v})}{\partial v_{ij} \partial v_{kl}} = 0$.

(ii) If $k = i$ and $l = j$, then

$$\frac{\partial^2 h(\mathbf{v})}{\partial v_{ij} \partial v_{kl}} = \frac{\partial^2 h(\mathbf{v})}{\partial v_{ij}^2} = \frac{1}{v_{ij}} - \frac{1}{\sum_{j=1}^N v_{ij}}.$$

(iii) If $k = i$ and $l \neq j$, then

$$\frac{\partial^2 h(\mathbf{v})}{\partial v_{ij} \partial v_{kl}} = \frac{\partial^2 h(\mathbf{v})}{\partial v_{ij} \partial v_{il}} = -\frac{1}{\sum_{j=1}^N v_{ij}}.$$

All these terms evaluated at $\boldsymbol{\pi}$ yield $\nabla^2 h(\boldsymbol{\pi})$, which will play a crucial role in approximating $D(\mathbf{\Gamma}_n \| \boldsymbol{\pi})$.

From the ergodicity of the chain \mathbf{Z} , it follows $\mathbf{\Gamma}_n \xrightarrow[n \rightarrow \infty]{a.s.} \boldsymbol{\pi}$.

It is seen that both $\nabla h(\mathbf{v})$ and $\nabla^2 h(\mathbf{v})$ are continuous in a neighborhood of $\boldsymbol{\pi}$, and we can use the second-order Taylor's series expansion of $h(\mathbf{v})$ centered at $\boldsymbol{\pi}$ to approximate $D(\mathbf{\Gamma}_n \| \boldsymbol{\pi}) = h(\mathbf{\Gamma}_n) - h(\boldsymbol{\pi})$. To be specific, we have

$$D(\mathbf{\Gamma}_n \| \boldsymbol{\pi}) \approx \frac{1}{2} (\mathbf{\Gamma}_n - \boldsymbol{\pi})' \nabla^2 h(\boldsymbol{\pi}) (\mathbf{\Gamma}_n - \boldsymbol{\pi}). \quad (19)$$

C. Threshold approximation

We use an empirical CDF to approximate the actual CDF of $D(\mathbf{\Gamma}_n \| \boldsymbol{\pi})$. In particular, by (19) we have

$$D(\mathbf{\Gamma}_n \| \boldsymbol{\pi}) \approx \frac{1}{2n} (\sqrt{n}(\mathbf{\Gamma}_n - \boldsymbol{\pi}))' \nabla^2 h(\boldsymbol{\pi}) (\sqrt{n}(\mathbf{\Gamma}_n - \boldsymbol{\pi})), \quad (20)$$

where, by Thm. III.3, $\sqrt{n}(\mathbf{\Gamma}_n - \boldsymbol{\pi}) \xrightarrow[n \rightarrow \infty]{d.} \mathcal{N}(0, \mathbf{\Lambda})$ with $\mathbf{\Lambda}$ given by (15). Thus, to derive an empirical CDF, we can generate a set of Gaussian sample vectors independently according to $\mathcal{N}(0, \mathbf{\Lambda})$ and then plug each such sample vector into (20) (i.e., replace $\sqrt{n}(\mathbf{\Gamma}_n - \boldsymbol{\pi})$) and compute the right hand side of (20), thus, obtaining a set of sample scalars, as an approximation for samples of $D(\mathbf{\Gamma}_n \| \boldsymbol{\pi})$.

Once we obtain an empirical CDF of $D(\mathbf{\Gamma}_n \| \boldsymbol{\pi})$, say, denoted $F_{em}(\cdot; n)$, then, by (5), we can use it to estimate η_n as

$$\eta_n^{wc} \approx F_{em}^{-1}(1 - \beta_n; n), \quad (21)$$

where $F_{em}^{-1}(\cdot; n)$ is the inverse of $F_{em}(\cdot; n)$.

Clearly, to calculate the right hand side of (20), we need the parameter $\boldsymbol{\pi}$. In practice, we typically do not know the actual values of the underlying chain parameters, so we need to estimate them accordingly. This can be done by computing $\mathbf{\Gamma}_n$ over a long past sample path.

IV. EXPERIMENTAL RESULTS

A. Numerical results for threshold approximation

The experiments in this subsection are performed using the software package TAHTMA [17]. For convenience, we use $\Theta = \{1, 2, \dots, N^2\}$ to indicate the states, and assume $\boldsymbol{\pi}$ (ground truth) is the initial distribution.

In the following numerical examples, we first randomly create a valid (i.e., such that (C) holds) $N \times N$ transition matrix \mathbf{Q} , hence an $N^2 \times N^2$ transition matrix \mathbf{P} , and then

generate T samples of the chain \mathbf{Z} , each with length n , denoted $\mathbf{Z}^{(t)} = \{Z_1^{(t)}, \dots, Z_n^{(t)}\}$, $t = 1, \dots, T$. These samples will be used to derive empirical CDF's. Also, to estimate the parameters of interest, we generate one more training sample $\mathbf{Z}^{(0)} = \{Z_1^{(0)}, \dots, Z_{n_0}^{(0)}\}$, where n_0 is the length. The stationary distribution $\boldsymbol{\pi}$ (estimation) of the chain is computed by taking any row of \mathbf{P}^{m_0} , where m_0 is a sufficiently large integer.

With samples $\mathbf{Z}^{(t)} = \{Z_1^{(t)}, \dots, Z_n^{(t)}\}$, $t = 1, \dots, T$, we can compute T samples of the scalar random variable $D(\boldsymbol{\Gamma}_n \| \boldsymbol{\pi})$, by (2). Thus, an empirical CDF of $D(\boldsymbol{\Gamma}_n \| \boldsymbol{\pi})$, denoted $F(\cdot; n)$, can be derived. We will treat $F(\cdot; n)$ as the actual CDF of $D(\boldsymbol{\Gamma}_n \| \boldsymbol{\pi})$. The threshold given by (21) with $F(\cdot; n)$ (i.e., $F(\cdot; n)$ plays the role of $F_{em}(\cdot; n)$) is then taken as the theoretical (actual) value η_n .

Next we derive an estimated empirical CDF of $D(\boldsymbol{\Gamma}_n \| \boldsymbol{\pi})$, using Taylor's series expansion together with Thm. III.3.

Let us first estimate the parameters of interest. Recall $\boldsymbol{\pi} = (\pi_{ij}; i, j = 1, \dots, N) = (\hat{\pi}_k; k = 1, \dots, N^2)$. We estimate $\hat{\pi}_k$ by

$$\hat{\pi}_k = \max \left\{ \frac{\sum_{i=1}^{n_0} \mathbf{1}_{\{Z_i^{(0)}=k\}}}{n_0}, \varepsilon \right\}, \quad k = 1, \dots, N^2, \quad (22)$$

where $\varepsilon > 0$ is a small number. The purpose of introducing ε is to avoid division by zero. Then, we obtain an estimated $\boldsymbol{\pi}$ as $\hat{\boldsymbol{\pi}} = (\hat{\pi}_k / \sum_{j=1}^{N^2} \hat{\pi}_j; k = 1, \dots, N^2)$, or written $\hat{\boldsymbol{\pi}} = (\hat{\pi}_{ij}; i, j = 1, \dots, N)$, where $\hat{\pi}_{11} = \hat{\pi}_1 / \hat{s}$, ..., $\hat{\pi}_{1N} = \hat{\pi}_N / \hat{s}$, ..., $\hat{\pi}_{N1} = \hat{\pi}_{(N-1)N+1} / \hat{s}$, ..., $\hat{\pi}_{NN} = \hat{\pi}_{N^2} / \hat{s}$, with $\hat{s} = \sum_{j=1}^{N^2} \hat{\pi}_j$ being the normalization constant.

We are now ready to calculate the estimate for $\nabla^2(\boldsymbol{\pi})$, by evaluating $\nabla^2(\hat{\boldsymbol{\pi}})$.

Further, by Lemma III.1, we estimate \mathbf{Q} by

$$\hat{q}_{ij} = \frac{\hat{\pi}_{ij}}{\sum_{t=1}^N \hat{\pi}_{it}}, \quad i, j = 1, \dots, N. \quad (23)$$

Then an estimation of \mathbf{P} , denoted $\hat{\mathbf{P}}$, can be directly derived from $\hat{\mathbf{Q}}$.

Finally, by (15), we estimate the covariance matrix $\boldsymbol{\Lambda}$ using

$$\hat{\Lambda}^{(i,j)} = \hat{\pi}_i (\mathbf{I}_{ij} - \hat{\pi}_j) + \sum_{m=1}^{m_0} [\hat{\pi}_i (\hat{\mathbf{P}}_{ij}^m - \hat{\pi}_j) + \hat{\pi}_j (\hat{\mathbf{P}}_{ji}^m - \hat{\pi}_i)], \quad i, j = 1, \dots, N^2. \quad (24)$$

It is worth noting that, to ensure the estimated $\boldsymbol{\Lambda}$ is at least symmetric, we update $\hat{\boldsymbol{\Lambda}}$ by $(\hat{\boldsymbol{\Lambda}} + \hat{\boldsymbol{\Lambda}}')/2$. Also, to ensure $\hat{\boldsymbol{\Lambda}}$ to be positive semi-definite, we use the trick of QR factorization [18]; for details, the reader is referred to [17].

We now generate T Gaussian sample vectors according to $\mathcal{N}(0, \hat{\boldsymbol{\Lambda}})$ and then plug each of them into (20) and compute the right hand side of (20) with $\nabla^2(\boldsymbol{\pi})$ replaced by $\nabla^2(\hat{\boldsymbol{\pi}})$, thus, obtaining T scalar samples as approximations for samples of $D(\boldsymbol{\Gamma}_n \| \boldsymbol{\pi})$. Therefore, an estimated empirical CDF of $D(\boldsymbol{\Gamma}_n \| \boldsymbol{\pi})$, denoted $F_{em}(\cdot; n)$, can be derived accordingly. η_n^{wc} given by (21) is then the WC approximation for the threshold.

Let the false alarm probability be $\beta_n = \beta = 0.001$. We set the aforementioned parameters as $N = 12$, $\varepsilon = 10^{-8}$, $T = 1000$, $m_0 = 1000$, and let $n_0 = 1000N^2$. In Fig. 1, the red line

indicates the theoretical value of the threshold η_n , the blue line shows the WC approximation η_n^{wc} , and the green line demonstrates the estimate η_n^{sv} obtained by Sanov's theorem, all as a function of the number of samples n . It is seen that η_n^{wc} is much more accurate than η_n^{sv} .

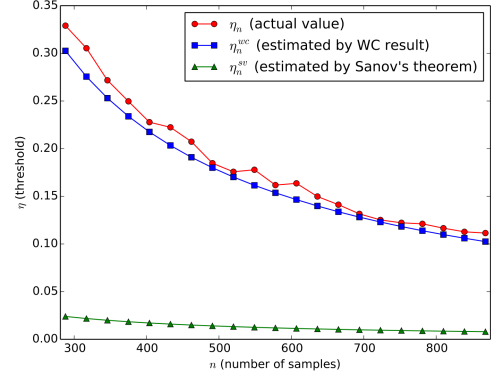


Fig. 1. Threshold (η) versus number of samples (n); the number of states in the original Markov chain is $N = 12$.

Remark 3: In Fig. 1, the red line corresponding to the actual value η_n is not as smooth; the reason is that each time when varying the number of samples n , we regenerate all the samples $\mathbf{Z}^{(t)} = \{Z_1^{(t)}, \dots, Z_n^{(t)}\}$, $t = 1, \dots, T$, from scratch. On the other hand, the blue line corresponding to η_n^{wc} looks smooth; this is because we only need to generate the T Gaussian sample vectors according to $\mathcal{N}(0, \hat{\boldsymbol{\Lambda}})$ once. Note that, in our experiments, most of the running time is spent generating the samples $\mathbf{Z}^{(t)} = \{Z_1^{(t)}, \dots, Z_n^{(t)}\}$, $t = 1, \dots, T$, and then calculating η_n . In practice, when implementing the approximation procedure proposed in this subsection, we will only need to focus on obtaining η_n^{wc} , which is computationally inexpensive.

B. Simulation results for network anomaly detection

We will use the term *traffic* and *flow* interchangeably in this subsection; they mean the same thing. The simulations are done using the software package SADIT [11], which, based on the *fs-simulator* [19], can efficiently generate flow-level network traffic datasets with annotated anomalies.

The network (see Fig. 2) consists of an internal network involving 8 normal users (*CT1-CT8*) and 1 server (*SRV*) that stores some sensitive information, and 3 Internet nodes (*INT1-INT3*) that visit the internal network via a gateway (*GATEWAY*).

When dealing with the data, we use the features “flow duration,” “flow size,” and “distance to cluster center” (this is related to a user IP address). The data preprocessing involves three steps: i) *network traffic representation and flow aggregation*, ii) *quantization*, and iii) *window aggregation*; cf. [4, Sec. III.A].

To implement Hoeffding's test, we first estimate a PL (resp., a set of PLs) for the stationary (resp., time-varying) normal traffic. For the former case, we use as the reference traffic the whole flow sequence with anomalies at some

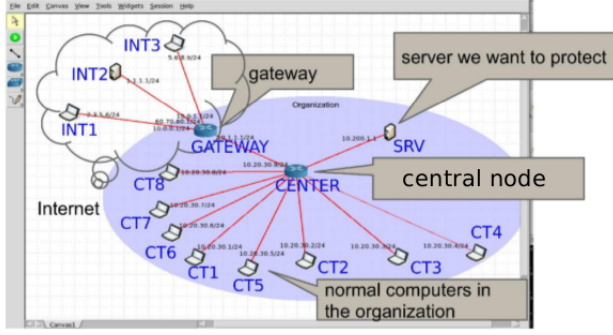


Fig. 2. Simulation setting (this figure is from [4] (or, see [20])).

time interval that we seek to detect. For the latter case, we generate the reference traffic and the traffic with anomalies separately, sharing all of the parameters except the ones regarding the anomalies. Using the windowing technique [20], [4], we then persistently monitor the traffic and report anomalies instantly whenever the relative entropy is higher than the threshold corresponding to the detection window.

The key difference between the whole detection procedures for the two types of traffic is that, estimating a PL for the stationary traffic is straightforward, while, for the time-varying traffic, we need to make an effort to identify several different PLs corresponding to certain time intervals. We apply the two-steps procedure proposed in [4]; that is, with some rough estimation of the shifting and periodic parameters of the traffic, we first generate a large set of PLs, and then refine the large family of PLs by solving a weighted set cover problem. For further theoretical details and the implementation, the reader is referred to [4], [11].

Note that, to deal with the time-varying traffic, [4] proposes a generalized Hoeffding's test, which we will use for Scenario 2 in the following. However, there is no essential difference compared to the typical Hoeffding's test. Note also that, in the following experiments, we only use the *model-based* method [20], [4].

1) *Scenario 1 (Stationary Network Traffic)*: We mimic the case for anomaly caused by "large access rate," cf. [20, Sec. IV.A.3]. The simulation time is 7000 s. A user suspiciously increases its access rate to 10 times of its normal value between 4000 s and 4500 s. The interval between the starting point of two consecutive time windows is taken as 50 s, the window-size is chosen to be $w_s = 400$ s, and the false alarm probability is set as $\beta = 0.001$.

Set the quantization level for "flow duration," "flow size," and "distance to cluster center" to be $n_1 = 1$, $n_2 = 2$, and $n_3 = 2$, respectively. Set the number of user clusters as $k = 3$.

See Fig. 3 and Fig. 4 for the detection results. Both figures depict the relative entropy (divergence) metric in (2). The green dashed line in Fig. 3 is the threshold calculated using Sanov's theorem (i.e., η_n^{sv} , where n is the number of flows in a specific detection window). The green dashed line in Fig. 4 is the threshold computed using the WC result derived in Sec. III (i.e., η_n^{wc}). The interval during which the entropy curve is above the threshold line (the red part) is the interval

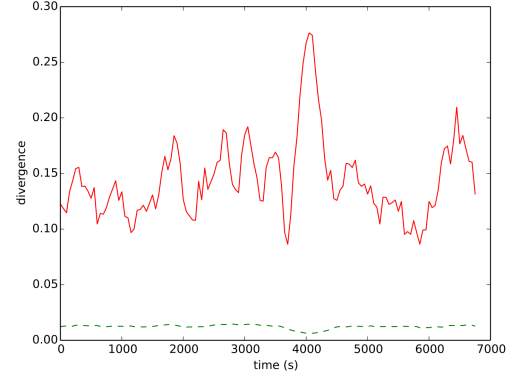


Fig. 3. Detection result for Scenario 1 with $w_s = 400$ s, $n_1 = 1$, $n_2 = 2$, $n_3 = 2$, $k = 3$; threshold is estimated by use of Sanov's theorem.

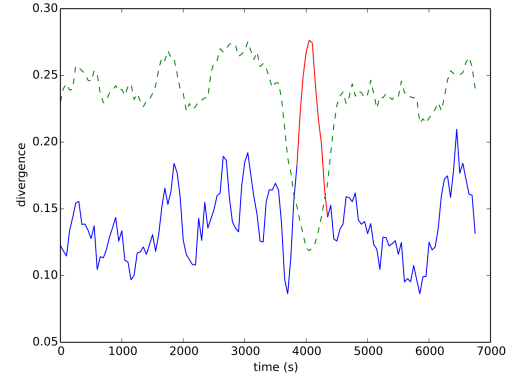


Fig. 4. Detection result for Scenario 1 with $w_s = 400$ s, $n_1 = 1$, $n_2 = 2$, $n_3 = 2$, $k = 3$; threshold is estimated by use of weak convergence result.

reported as abnormal.

Fig. 3 shows that, if using η_n^{sv} as the threshold, then the detection method reports an anomaly for all windows including the ones wherein the behavior of the network traffic is actually normal; in other words, there are too many false alarms. Clearly, this is undesirable. Fig. 4 shows that, if, instead, we use η_n^{wc} as the threshold, then the detection method will not report any false alarm while successfully identifying the true anomalies between 4000 s and 4500 s. The reason for such detection results is that η_n^{wc} is more accurate than η_n^{sv} .

2) *Scenario 2 (Time-Varying Network Traffic)*: Consider the network with day-night pattern where the flow size follows a log-normal distribution. We use exactly the same scenario as that in [4, Sec. IV.B.2].

Applying the same procedure as in [4, Sec. III.C], we first obtain 32 rough PLs. Then, using the heuristic PL refinement algorithm given in [4, Sec. III.D] equipped with the up-bound nominal cross-entropy parameter $\lambda = 0.027565$, which is determined by the WC approximation, we finally obtain 6 PLs (see Fig. 5). These PLs are active during morning, afternoon, evening, mid-night, dawn, and the transition time around sunrise, respectively. In the following detection pro-

cedure, the key difference between our method and the one used in [4] is that we no longer manually set the threshold universally as a constant; instead, the threshold η_n^{wc} for each detection window is automatically calculated by use of the WC approximation. We set the quantization level for “flow duration,” “flow size,” and “distance to cluster center” to be $n_1 = 1$, $n_2 = 4$, and $n_3 = 1$, respectively, and set the number of user clusters to $k = 1$. The interval between the starting point of two consecutive time windows is chosen as 1000 s, the window-size is taken as $w_s = 1000$ s, and the false alarm probability is set as $\beta = 0.001$.

Noting that the anomaly is simulated beginning at 59 h and lasting for 80 minutes [4], we see from Fig. 6 that the anomaly is successfully detected, without any false alarms.

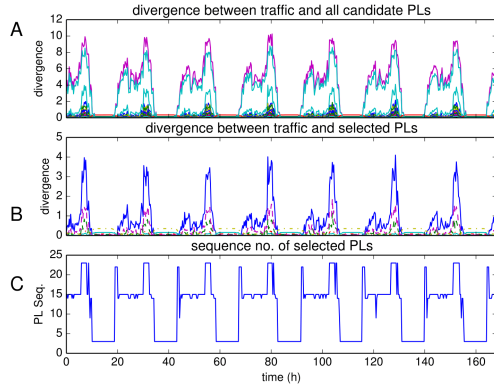


Fig. 5. PLs identification for Scenario 2 with $w_s = 1000$ s, $n_1 = 1$, $n_2 = 4$, $n_3 = 1$, $k = 1$.

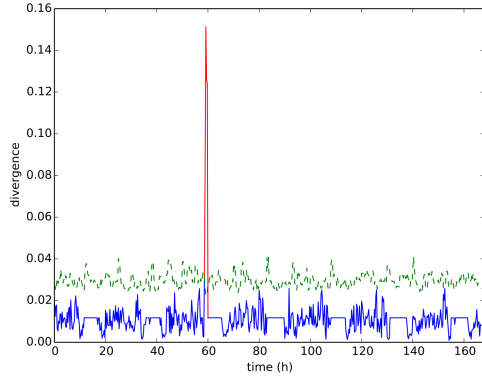


Fig. 6. Detection result for Scenario 2 with $w_s = 1000$ s, $n_1 = 1$, $n_2 = 4$, $n_3 = 1$, $k = 1$.

V. CONCLUSIONS

We establish a weak convergence result for the relative entropy in Hoeffding’s test under a Markovian assumption, which enables us to obtain a more accurate approximation (compared to the existing approximation given by Sanov’s theorem) for the threshold needed by the test. We validate the accuracy of such approximation through numerical experiments and simulations for network anomaly detection.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Jing Wang for his contributions and help in developing the software package SADIT [11].

REFERENCES

- [1] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*. Springer, 1998.
- [2] I. C. Paschalidis and G. Smaragdakis, “Spatio-temporal network anomaly detection by assessing deviations of empirical measures,” *IEEE/ACM Transactions on Networking (TON)*, vol. 17, no. 3, pp. 685–697, 2009.
- [3] S. Meyn, A. Surana, Y. Lin, and S. Narayanan, “Anomaly detection using projective Markov models in a distributed sensor network,” in *Proceedings of the 48th IEEE Conference on Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009.*, Dec 2009, pp. 4662–4669.
- [4] J. Wang and I. Paschalidis, “Statistical traffic anomaly detection in time-varying communication networks,” *IEEE Transactions on Control of Network Systems*, vol. 2, no. 2, pp. 100–111, 2015.
- [5] J. Unnikrishnan and D. Huang, “Weak convergence analysis of asymptotically optimal hypothesis tests,” *IEEE Transactions on Information Theory*, submitted, 2013.
- [6] W. Hoeffding, “Asymptotically optimal tests for multinomial distributions,” *The Annals of Mathematical Statistics*, pp. 369–401, 1965.
- [7] J. Zhang, “Threshold Approximation for Hoeffding’s Test under I.I.D. Assumption (TAHTIID),” <https://github.com/jingzbu/TAHTIID>, 2015.
- [8] J. Unnikrishnan, D. Huang, S. P. Meyn, A. Surana, and V. V. Veeravalli, “Universal and composite hypothesis testing via mismatched divergence,” *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1587–1603, 2011.
- [9] S. S. Wilks, “The large-sample distribution of the likelihood ratio for testing composite hypotheses,” *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938.
- [10] M. Iltis, “Sharp asymptotics of large deviations in \mathbb{R}^d ,” *Journal of Theoretical Probability*, vol. 8, no. 3, pp. 501–522, 1995.
- [11] J. Wang, J. Zhang, and I. C. Paschalidis, “Systematic Anomaly Detection of Internet Traffic (SADIT),” <https://github.com/hbhzwj/SADIT>, 2014.
- [12] G. L. Jones, “On the Markov chain central limit theorem,” *Probability Surveys*, vol. 1, pp. 299–320, 2004.
- [13] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. New York: Springer, 1993.
- [14] G. O. Roberts and J. S. Rosenthal, “General state space Markov chains and MCMC algorithms,” *Probability Surveys*, vol. 1, pp. 20–71, 2004.
- [15] F. Pène, “Rate of convergence in the multidimensional central limit theorem for stationary processes. Application to the Knudsen gas and to the Sinai billiard,” *The Annals of Applied Probability*, vol. 15, no. 4, pp. 2331–2392, 2005.
- [16] W. Feller, *An introduction to probability theory and its applications*. John Wiley, 1957, vol. 2.
- [17] J. Zhang, “Threshold Approximation for Hoeffding’s Test under Markovian Assumption (TAHTMA),” <https://github.com/jingzbu/TAHTMA>, 2015.
- [18] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.
- [19] J. Sommers, R. Bowden, B. Eriksson, P. Barford, M. Roughan, and N. Duffield, “Efficient network-wide flow record generation,” in *INFOCOM, 2011 Proceedings IEEE*. IEEE, 2011, pp. 2363–2371.
- [20] J. Wang, D. Rossell, C. G. Cassandras, and I. C. Paschalidis, “Network anomaly detection: A survey and comparative analysis of stochastic and deterministic methods,” in *IEEE 52nd Annual Conference on Decision and Control (CDC)*, Dec 2013, pp. 182–187.